Mechanistic Model Development for a Multi-Enzyme Protein Hydrolysis Process

by

Tito Nilsson

Department of Chemical Engineering Lund University,
in collaboration with Biofynt ApS, Copenhagen

June 2023

Supervisor: **Ph.D Niklas Andersson**
Co-supervisor: **Ph.D Paloma Santacoloma**
Examiner: **Professor Bernt Nilsson**

Picture on front page:

Postal address
Box 124
SE-221 00 Lund, Sweden
Web address
http://www.lth.se/chemeng/

Visiting address
Kemicentrum
Naturvetarvägen 14
223 62 Lund, Sweden

Telephone
+46 46-222 82 85
+46 46-222 00 00

Acknowledgement

This thesis was conducted at the Department of Chemical Engineering at Lund University in collaboration with Biofynt ApS, Copenhagen, and concludes my education at the Chemical Engineering Program at Lund University.

I would like to express gratitude towards Paloma Santacoloma, Angela Ruales Salcedo, and Luna Santacoloma at Biofynt for providing the opportunity for this thesis. The project provided an interesting and challenging task from which I have gained significant personal development. I would like to express an extra “thank you” to Paloma for taking time each week to have interesting discussions and to guide me along the project. And, for keeping my motivation high with seemingly endless optimism.

I would also like to thank Bernt Nilsson for examining this project, and Niklas Andersson for providing valuable input and guidance as supervisor. Further I would like to thank both Bernt and Niklas for providing courses during my master years that gave me essential knowledge for writing this thesis.

Finally, I would like to say ‘thank you’ to all the people in the Symbion coworking space. Who provided a fun and interesting environment to conduct this thesis in.

Thank you!

Abstract

New food products are constantly put on the market. Today the focus often lies in creating nutritional and sustainable products to reduce the environmental impact of the product in question. One example of this is the approval of the yellow mealworm, *Tenebrio Molitor*, as a novel food in the European union. To efficiently include the protein fraction of the yellow mealworm into food formulations, knowledge of the techno-functional properties of the protein is crucial. The techno-functional properties of the protein have further been demonstrated to be manipulable by partial enzymatic hydrolysis. This thesis aimed to develop a mechanistic mathematical model that describes the change in the molecular weight distribution of the protein due to multi-enzymatic hydrolysis. Which could later be used to tailor the techno-functional properties. The developed model showed good result in being able to capture the expected multi-enzyme systems behavior, but it was necessary to introduce empirical adjustment factors to calibrate the model adequately with literature data. Synthetic data was created to examine the potential of accurately estimating the model parameters and to examine the temperature and pH dependence of the system through response surface methodology. Estimation of the parameters with the synthetic data revealed that the optimization procedure is likely to be sensitive to converge in local minimums. Correlation matrices also revealed a high interdependence among the parameters, making the task of finding a unique optimal calibration difficult. The empirical model used in the response surface method showed potential to capture the expected temperature and pH dependencies of the model, if given correct estimations.

Sammanfattning

Nya livsmedel dyker hela tiden upp på marknaden. Ofta ligger fokus på att skapa hälsosamma och miljövänliga produkter som minskar vår påverkan på klimatet. Ett exempel på en detta är godkännandet av mjölbaggen, *Tenebrio Molitor*, som ett livsmedel i Europeiska unionen. För att effektiv inkludera proteinet från mjölbaggen i livsmedelsprodukter, är kunskap om proteinets funktionella egenskaper avgörande. De funktionella egenskaperna har påvisats vara föränderliga genom hydrolys med hjälp av enzymer. Denna tes syftade till att formulera en mekanistisk matematisk modell som beskriver förändringen i molekylviktsfördelningen hos proteinet när det hydrolyseras av olika enzymer. Denna modell kan sedan användas för att skräddarsy proteinets funktionella egenskaper till olika livsmedel. Modellen visade potential att kunna fånga det förväntade beteendet i hydrolyssystemet, men tillägget av empiriska parametrar var nödvändigt för att få tillräckligt bra kalibreringsresultat gentemot data i tillgänglig litteratur. Syntetiska dataset skapades för att undersöka möjligheterna att exakt uppskatta värdet hos modellparametrarna och även för att fånga temperatur- och pH-beroendet i processen genom ”Ytresponsmetodik”. Resultatet från parameterkalibreringar med den syntetiska datan påvisade att optimeringsproceduren sannolikt är känslig för att fastna i lokala minima. Vidare beräknades korrelationsmatriser för parameterskattningarna som visade ett starkt beroende mellan parametrarna, vilket innebär att det är svårt att hitta en unik optimal skattning. Temperatur- och pH-beroendet hos modellen kan sannolikt fångas med hjälp av den förslagna empiriska modellen inom ytresponsmetodiken, men kräver i så fall korrekta parameterskattningar.

Popular Science Summary

Most people have probably noted the recent upswing in new food products that are promoted as being more environmentally friendly to consume. One recent development is the start of research regarding insects as food. Although insect products have not reached the store shelves yet to a great extent, plenty of research is ongoing for large scale production of food products containing insects. One insect that recently was approved in the European union as a novel food is the yellow mealworm *Tenebrio Molitor*. The protein in the yellow mealworm is nutritious and could have health benefits for human if it was included in food products, while being more environmentally friendly compared to other sources of protein.

Proteins are long chains of similar molecules that are called amino acids. The length of the chains can vary drastically, from one or two, up to several hundreds, or more. When the protein is extracted from its source, the mixture will contain proteins with very different lengths. The ratio between the amounts of the different protein lengths will have a big impact on the food product that it is included in. In terms of taste, texture etc. The impact on the food product can be controlled by splitting the proteins into shorter chains before introducing it into the food product. This can be done with the help of enzymes and is called hydrolysis.

In this thesis, a model was constructed to describe the hydrolysis of proteins with the help of enzymes. A model is essentially math that describes what would happen in a real experiment. Models are never perfect, but if they are good enough, they can give you a good idea of what happens and can save much time and resources since less lab work is required. The created model showed potential to be useful in predicting how the different chain lengths would change due to the enzymatic hydrolysis. But it was also shown that determining the exact mathematical expressions of the model can be difficult and more work is required.

Populärvetenskaplig Sammanfattning

Fler och fler livsmedelsprodukter dyker upp på marknaden som marknadsförs som miljövänliga, jämfört med liknande alternativ. På senaste tiden har forskning om insekter som livsmedel fått ett lyft, även om få produkter finns tillgängliga för konsumenter än. Mjölbaggen, *Tenebrio Molitor*, är en insekt som nyligen blev godkänt livsmedel i Europeiska unionen. Proteinet i mjölbaggar anses näringsrikt och kan vara ett hälsosamt tillägg i livsmedelsprodukter. Samtidigt som det är mer miljövänligt än många andra proteinkällor.

Proteiner är långa kedjor av molekyler som kallas aminosyror. Kedjornas längd kan variera drastisk, från enstaka syror till flera hundra, eller fler. När protein extraheras från källan kommer extraktionen innehålla en blandning av alla möjliga kedjelängder. Förhållandet mellan andelarna av olika längder har stor påverkan på smak och konsistens när det inkluderas i livsmedel. Påverkan som protein har kan justeras genom att dela upp proteinkedjorna till kortare kedjor med hjälp av enzymer. Denna process kallas hydrolys.

Den här tesen syftade till att skapa en modell som beskriver hydrolysen av protein med hjälp av olika enzymer. En modell är, enkelt sagt, matematik som beskriver vad som hade hänt i ett verkligt experiment. Modeller beskriver aldrig verkligheten exakt men om dem gör ett tillräckligt bra jobb kan de ge bra uppskattningar. Detta kan i sin tur spara mycket tid och resurser då färre verkliga experiment behöver utföras. Modellen som skapades visade god potential att beskriva hydrolysprocessen och hur kedjelängderna varierar på grund av enzymerna. Men samtidigt påvisades det att ta fram de exakta matematiska uttrycken i modellen kan vara svårt, och vidare utveckling av modellen är nödvändigt.

Nomenclature

Roman Letters

Superscript "en" and "ex" refers to endopeptidase and exopeptidase respectively

A	*Pre-exponential factor*	*"Same as rate constant"*
c_p	*Total peptide concentration*	$[\ mmol\ L^{-1}\ ]$
E_A	*Activation energy*	$[\ \mathrm{J\ mol^{-1}}]$
E_f	*Free enzyme concentration*	$[\ g\ enzyme\ L^{-1}\ ]$
$\langle E+P \rangle$	*Enzyme-Substrate complex*	$[\ g\ enzyme\ L^{-1}\ ]$
E_t	*Total enzyme concentration*	$[\ g\ enzyme\ L^{-1}\ ]$
F_{AA}	*Fraction of specific amino acid in the protein*	
H	*Concentration of hydrolyzed peptide bonds*	$[\ mmol\ bonds\ L^{-1}\ ]$
$J(\theta)$	*Cost function as the sum of square errors*	
k	*Reaction rate constant*	$[\ mmol\ g^{-1}\ enzyme\ min^{-1}]$
K_{dis}	*Dissociation constant*	$[\ mmol\ bonds\ L^{-1}\ ]$
K_I	*Inhibition constant*	$[\ mmol\ L^{-1}\ ]$
M	*Molecular weight*	$[\ g\ mol^{-1}\ ]$
N	*Maximal degree of polymerization*	
N	*Total number of datapoints*	
N_{DS}, N_y, N_t	*Number of datasets, model outputs, and time evaluations respectively*	
P_i	*Concentration of peptide with length i*	$[\ mmol\ L^{-1}\ ]$
R	*Universal gas constant*	$[\ \mathrm{J\ K^{-1}\ mol^{-1}}]$
r_{P_i}	*Net reaction rate of* P_i	$[\ mmol\ L^{-1}\ min^{-1}\ ]$
$RMSD$	*Root-mean-square deviation*	

S_0	*Total protein concentration*	$[\ g\ protein\ L^{-1}\ ]$
s^2	*Error of variance*	
T	*Temperature*	$[K]\ or\ [℃]$
t	*Time*	[min]
u	*Vector of model inputs*	
x	*Vector of state-variables*	
X	*Model matrix is RSM*	
y	*Model output*	
$\hat{y}$	*Datapoint*	
y_0	*Initial Values of lumped components*	$[\ g\ protein\ L^{-1}\ ]$
Y	*Response vector in RSM*	

Greek Letters

$\boldsymbol{\beta}$	*Vector of coefficients in RSM*
$\boldsymbol{\varepsilon}$	*Vector of errors in RSM*
ε^{en}	*Fraction of hydrolysable peptide bonds in the initial protein for endopeptidase*
ε^{ex}	*Fraction of hydrolysable peptide bonds in the initial protein for exopeptidase*
$\boldsymbol{\theta}$	*Vector of parameters*
θ_{min}	*Lower bound for parameter estimation*
θ_{max}	*Upper bound for parameter estimation*
ξ	*Adjustment parameters*
ϕ^{en}	*Fraction of hydrolysable bonds for endopeptidase*
ϕ^{en}	*Fraction of hydrolysable bonds for exopeptidase*

Table of Content

1 Introduction

1.1 Overview

With increasing focus on the environmental impact of the human population's consumption, many new food products have been put on the market in recent years. These products often market the environmental benefits compared to other products, in terms of greenhouse emissions, water usage, land usage etc. One recent development in this area is the placing of the yellow mealworm (*Tenebrio Molitor*) on the European market as a novel food. Biofynt Aps is a recently founded Danish start-up that explores the possibilities of effectively including yellow mealworm protein in food formulations.

An important concept for introducing protein into food formulations is the techno-functional properties of the protein e.g., viscosity, emulsifying activity, and oil-holding capacity. These are crucial properties for effective food formulations and have been shown to be manipulable through enzymatic hydrolysis of protein.

This thesis project constitutes initial efforts at Biofynt to develop a mechanistic mathematical model that would describe the change in molecular weight distribution of the extracted mealworm protein as it undergoes hydrolysis in a multi-enzyme system. The long-term goal (not explored in this thesis) is to link the state of the protein weight distribution to several techno-functional properties, to improve product design. The complexity of the system creates a difficult modeling scenario, and the calibrated model will rely heavily on the data used. Limited data on the system in question exist in literature and practical limitations prevented the experimental work required to acquire relevant data. The focus of this thesis was therefore not to propose a validated model, but rather to propose a structure that could serve as a basis for both process understanding and further model development for practical implementation in the future. Synthetic data was generated based on the limited data available in the literature to implement the programming structure for parameter estimation and identifiability. The synthetic data was then utilized to calibrate the model, and these results were used to evaluate the potential of the model structure, in terms of prediction accuracy and parameter

identifiability. Additionally, a way of including temperature and pH as model inputs was briefly explored by response surface methodology.

1.2 Aims

The aims of the thesis are stated below.

- Research literature regarding modeling enzymatic hydrolysis of protein and propose a mechanistic model that captures the change in molecular weight distribution of the protein to better understand the reaction and process dynamics.
- Generate synthetic data to implement the programming structure for parameter estimation and identifiability. And to examine opportunities and issues of future model calibration.
- To evaluate strengths and weaknesses regarding the structure of the model.
- Explore the inclusion of temperature and pH as input variables in the model.

1.3 Outline

The general structure of this thesis is summarized below.

- Chapter 2 provides the reader with relevant background information and motivates as to why this thesis holds relevance and could provide useful insights into the development of a new food product.
- Chapter 3 goes through the necessary theoretical information that is useful in establishing the mathematical model.
- Chapter 4 discusses the mathematical derivation of the model and the assumptions made. It continues with methods used in optimization, statistical computation, and response surface methodology.
- Chapter 5 presents and discusses the results.
- Chapter 6 summarizes the key findings of the thesis project and suggests strategies for future development of the model.

2 Background

2.1 Insects as a Future Food

Although the consumption of insects has had a long running place in human history, it is only recently that the interest in producing insects at the industrial scale for a global food source has sparked. The interest is likely the result of the potential nutritional and environmental advantages of insects compared to other food sources that have been proposed (Huis et al., 2013). Insects could be a valuable source of protein intake in the human diet, where dry insect matter has shown average protein contents between 35% and 61% and with all essential amino acids. Further, insect production has been shown to likely be a more sustainable and environmentally friendly way of producing nutrients, compared to other animals, requiring less water, having a better feed-to-meat conversion and emitting less greenhouse gas (Lange and Nakamura, 2021).

2.2 Food Applications and Techno-Functional Properties

The development of insect protein as part of human consumption requires much research on the topic. To date, there is only a limited amount of research regarding the techno-functional properties of insect protein. Knowledge of the techno-functional properties and ways of manipulating them are essential for effective use of protein in food formulations (Gkinali et al., 2022b). Examples of techno-functional properties in food applications are solubility, emulsifying activity, oil-holding capacity, and foam capacity.

The degree of hydrolysis, which is a measure of the amount of peptide bonds in the protein that has been hydrolyzed, has been demonstrated to impact several techno-functional properties of insect protein when enzymes are utilized (Leni et al., 2020) (Purschke et al., 2018). Another characteristic of protein that has importance in food formulations is the molecular weight distribution, and the change of it due to hydrolysis. This has been studied in order to observe the degradation of allergens (García Arteaga et al., 2020), and the release of bioactive peptides (Rivero Pino et al., 2020). Further, Kristoffersen et al. (2020) concluded that the degree of hydrolysis and molecular weight distribution of protein hydrolysates offer complementary information and adequately describes the state of the hydrolysis process.

2.3 Biofynt ApS

Biofynt ApS (a B2B start-up founded 2022) is a Danish company that is developing a high protein powder ingredient to enable food manufacturers to enrich the nutritional value of daily consumed food products without compromising the appearance, taste, or texture. In this way, the end-consumer can enjoy a traditional product (e.g. pasta, bread, meatballs etc) that is tasty, nutritional and also sustainable.

Biofynt is exclusively focusing on mealworm, *Tenebrio molitor*, processing (e.g. extraction and purification of valuable fractions). The initial effort is on the protein fraction which is carefully extracted and functionalized using multi-enzyme technology to potentialize nutrition and sensorial experience in finished food and drink products.

3 Theory

3.1 T. Molitor Protein

The amino acid composition of the protein in yellow mealworm has been evaluated many times in literature, mainly from the perspective of nutritional value (Costa et al., 2020). In this thesis the profile of amino acids is used to compute an average molecular weight for the residues. The method is described in section 4.2.4. The amino acid composition according to Costa et al. (2020) is presented in table 3.1. Further, the water-soluble proteins in *T. Molitor* have previously been shown to mainly have molecular weights below 14 kDa (Gkinali et al., 2022a).

Table 3.1: Amino Acid content in T. molitor Protein (Costa et al., 2020).

Amino acid	*T. molitor* larvae (mg/g protein)
Histidine (HIS)	28,9
Isoleucine (ILE)	32,1
Leucine (LEU)	75,9
Lysine (LYS)	26,2
Threonine (THR)	54,8
Valine (VAL)	49,2
Aspartic acid (ASP)	97,4
Glutamic acid (GLU)	125,3
Serine (SER)	58,1
Glycine (GLY)	66,0
Arginine (ARG)	45,6
Alanine (ALA)	100,9
Tyrosine (TYR)	67,2
Phenylalanine (PHE)	34,3
Proline (PRO)	95,9
Total a.a.	957,8
Total e.a.a.	368,6

3.2 Enzymes

Using enzymes for the hydrolysis of food proteins has been developed as an alternative to hydrolysis by acids or alkali. Proteases, enzymes that catalyzes the cleavage of peptide bonds, have several advantages compared to the chemical methods. Such as lighter color and milder taste of the protein. Hydrolysates generated from enzymatic hydrolysis are also generally considered safe. Additionally, enzymes can potentially offer larger flexibility because of their

specificity, which enables the user to control the outcome of the hydrolysis process. Combined or sequential use of different proteases have recently gained attention because of this flexibility. (Habinshuti et al., 2023) Two popular commercial enzyme preparations are Alcalase and Flavourzyme, which are discussed in more detail below. The operating conditions for these enzyme preparations are presented in table 3.2 (Beaubier et al., 2021).

Table 3.2: Process condition domains for Alcalase and Flavourzyme.

Enzyme	*Temperature* (℃)		*pH*		*E/S*	
	Min	Max	Min	Max	Min	Max
Alcalase	45	75	7.0	10.0	1/150	1/15
Flavourzyme	40	60	5.0	8.5	1/150	1/15

3.2.1 Alcalase

Commercially known as Alcalase, this protease is obtained from the bacteria *Bacillus licheniformis* and is an alkaline endopeptidase. Meaning that it hydrolyses peptide bonds of non-terminal amino acids (Chew et al., 2019). It uses a catalytic triad (Asp32, His64, Ser221) for the hydrolysis of the peptide bonds (Graycar et al., 2013). Alcalase has been described to have a preference for large uncharged residues at the site of hydrolysis but it is generally considered to have broad specificity and always yield high degrees of hydrolysis (Tacias-Pascacio et al., 2020).

3.2.2 Flavourzyme

Flavourzyme is a commercial enzyme blend from Novozymes (Bagsvaerd, Denmark) that has been developed with a flavor altering properties in mind. Although it has been developed and is marketed towards flavor adjustments, it has been applied in several proteolysis applications with different purposes. Rivero Pino et al. (2020) investigated certain biological activities of *Tenebrio Molitor* protein hydrolysates. The Flavourzyme blend is derived from the fungus *Aspergillus oryzae* and comprises both endo- and exoproteases. However mostly the latter. Exoproteases are generally associated with hydrolysis close to the N-terminus of the peptide

chain. Research into what specific enzymes constitutes the mixture have previously been done and found eight key enzymes, which are presented in table 3.3 (Merz et al., 2015).

Table 3.3: Different Enzyme categories within the Flavourzyme Preparation.

Enzyme	**Enzyme Category**
Leucine Aminopeptidase A	Aminopeptidase
Leucine Aminopeptidase 2	Aminopeptidase
Dipeptidyl Peptidase 4	Dipeptidyl peptidases
Dipeptidyl Peptidase 5	Dipeptidyl peptidases
Neutral Protease 1	Endopeptidase
Neutral Protease 2	Endopeptidase
Alkaline Protease 1	Endopeptidase
Alpha-amylase A type 3	Amylase

3.3 Temperature and pH dependence

The enzymatic hydrolysis of protein is heavily impacted by the process conditions which the hydrolysis is performed under. The main conditions are temperature, pH, and concentrations of both protein substrate and enzyme. These conditions affect both the catalytic properties of the enzymes as well as the structure of protein. Consequently, they impact both the mechanism of the hydrolysis and the reaction kinetics (Beaubier et al., 2021). A common way of describing the effect of temperature on reaction rate constants (regardless of reaction order) is with the Arrhenius equation (equation 1). Where k is the rate constant, A is the pre-exponential factor, E_a is the activation energy, and R is the universal gas constant (Atkins et al., 2017).

$$k = A * e^{\frac{-E_A}{RT}} \quad (1)$$

Several studies have found that this relation well describes the effect of temperature on the reaction rate of enzymatic hydrolysis of protein (Qi and He, 2006) (Ruan et al., 2010) (Márquez

and Vázquez, 1999). However, the Arrhenius equation is limited to describing the effect of temperature on the reaction rate and other process conditions needs to remain constant when the parameters of equation 1 are estimated.

As stated above, pH has an impact on the kinetic parameters when modeling the kinetic behavior of enzymatic proteolysis. De Pretto et al. (2022) clearly showed the influence of pH (at values of 8, 9, and 10) on the hydrolysis of soybean meal protein with subtilisin. Where the degree of hydrolysis (amount of cleaved peptide bonds) significantly varied. The impact of pH on the degree of hydrolysis has also been shown on the hydrolysis of egg-white protein by pepsin (Ruan et al., 2010). Sousa et al. (2004) had great success fitting the rate constant, the Michaelis-Menten constant, and the inhibition constant to bell curves in the entire experimental pH spectrum (6 – 11). The kinetic parameters were obtained by kinetic fitting of data from the hydrolysis of whey protein by Alcalase.

A way of capturing the temperature and pH dependency of the model is through Response Surface Methodology (RSM). RSM provides a method to empirically derive the nature of the relationship between process conditions and the outcome of the process. Broadly speaking, RSM is not only the fitting of a surface response to independent variables, but a collection of techniques that can help with designing future experiments, optimization of the process, outlier identification, etc. (Sarabia and Ortiz, 2009). It should be noted that RSM does not rely on any mechanistic mathematical derivation and no theoretical knowledge of the process is included in the method. Therefore, the estimated coefficients lack physical meaning. But it is an simple and effective way of gaining insight into the process (Valencia et al., 2015). In this thesis, RSM is utilized to observe the impact of pH and temperature on the model parameters.

4 Methods in Model Construction and Evaluation

4.1 Model Objective

In adherence to section 2.2, the degree of hydrolysis and the molecular weight distribution of the hydrolysates are important factors to consider for food applications. These properties are therefore the main targets for the model output. “Simple” models have been developed to describe the degree of hydrolysis for protein by enzymatic degradation (Zhou et al., 2016) (Ruan et al., 2010). Predicting the molecular weight distribution have also been tried. Shi et al. (2005) developed a model for the hydrolysis of bovine serum albumin by trypsin. De Pretto (2022) utilized a neural network to predict the molecular weight distribution from the degree of hydrolysis of soybean meal protein with subtilisin. However, both these models output lumped components into “pseudo components” that include a spectrum of molecular weights.

To the extent of my knowledge, no mechanistic model that would allow for a complete molecular weight distribution as output has been developed for the enzymatic hydrolysis of protein. However, models that describe the enzymatic hydrolysis of cellulose including a generated complete molecular weight distribution have been derived (Zhang and Lynd, 2006) (Niu et al., 2016). The work in these articles provide the basis for the model developed in this chapter.

4.2 Model Assumptions

To reasonably describe the complex hydrolysis system in mathematical terms, several simplifying assumptions are made.

1) The reaction takes place in a well stirred tank. Hence, no consideration is given to spatial derivates, and concentrations, temperature, and pH is assumed to be homogenous throughout the system. Additionally, the process conditions are assumed to be kept at steady state by external control of pH and temperature. No variations in this, as a result from the hydrolysis reaction, is considered by the model.
2) The model reduces the commercial enzyme preparations into two distinct general enzymes. Endopeptidase and Exopeptidase.
3) The peptide chains are considered to have the same repeating unit, a “general” amino acid with the molecular weight calculated in section 4.2.4. Therefore, all bonds in the peptide chains are assumed to interact identically with the endopeptidase. This means

that when a peptide of length k is cleaved into two peptides of length i and j ($j \leq i$), the probability of the length i being formed from the length k is $\frac{2}{k-1}$ (Niu et al., 2016).

4) The exopeptidase is considered to only cleave the two bonds closest to the N-terminal of a peptide chain, with equal probability of each bond being cleaved. Hence, a peptide of length i can only be produced by exopeptidase reacting with a peptide of length k where $k = i + 1, i + 2,$ and the probability of length i being formed from k is $\frac{1}{2}$.
5) The quasi steady-state approximation holds for all complexes formed between peptides and enzymes.
6) Peptides of lengths between one and five are assumed to not contain any cleavable bonds and are not consumed by any reaction. This assumption seems reasonable given the specificity of the enzymes and the small amount of peptide bonds within the short chains.
7) Peptides with polymerization degrees less than 56 are assumed to inhibit both enzymes competitively.
8) Both enzymes are assumed to interact indistinguishably with all peptide lengths above five. Therefore, only one rate constant, one dissociation constant, and one inhibition constant for each enzyme is necessary to describe the system.
9) For the simulated process conditions and time spans, loss of enzyme activity is considered to be negligible. This assumption is based on the fact that conditions are relatively mild relative to the optimal conditions for the enzyme preparations.

4.3 Mathematical Derivation

4.3.1 Enzymatic Hydrolysis by Endopeptidase

The hydrolysis of non-terminal peptide bonds is described by modified Michaelis-Menten kinetics represented in the scheme below.

$$P_i + E_f^{en} \overset{K_{dis}^{en}}{\longleftrightarrow} \langle E^{en} + P_i \rangle \overset{k^{en}}{\longrightarrow} E_f^{en} + P_{i-j} + P_j$$

$$P + E_f^{en} \overset{K_I^{en}}{\longleftrightarrow} \langle E_I^{en} + P \rangle$$

Where P_i is a peptide chain with the length of i amino acids, E_f^{en} a free endopeptidase enzyme and $\langle E^{en} + P_i \rangle$ the complex of a peptide chain of length i bound to an enzyme. P is an inhibiting peptide chain and $\langle E_I^{en} + P \rangle$ the inhibited enzyme complex. K_{dis}^{en} and K_I^{en} are equilibrium constants and k^{en} the reaction rate constant for the enzyme complexed peptide chains.

The total number of peptide bonds, or substrate quantity is $\sum_{n=6}^{N}[(n-1)P_n]$, where N is the largest degree of polymerization. Therefore, the dissociation constant K_{dis}^{en} can be defined as,

$$K_{dis}^{en} = \frac{E_f^{en}\phi^{en}(i-1)P_i}{\langle E^{en} + P_i \rangle} \cong \frac{E_f^{en}\phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}{\sum_{n=6}^{N}[\langle E^{en} + P_n \rangle]} \tag{2}$$

where ϕ^{en} is the fraction of peptide bonds that are hydrolysable by the enzyme and is defined in section 4.2.4. Equation 2 can be rearranged into equation 3.

$$\langle E^{en} + P_i \rangle \cong \frac{(i-1)P_i}{\sum_{n=6}^{N}[(n-1)P_n]} \sum_{n=6}^{N}[\langle E^{en} + P_n \rangle] \tag{3}$$

The total amount of peptide is defined as $\sum_{n=1}^{N}[P_n]$, which gives the definition of K_I^{en}.

$$K_I^{en} = \frac{E_f^{en}P_i}{\langle E_I^{en} + P_i \rangle} \cong \frac{E_f^{en}\sum_{n=1}^{N}[P_n]}{\sum_{n=1}^{N}[\langle E_I^{en} + P_n \rangle]} \tag{4}$$

The total mass balance of endopeptidase enzyme is given by,

$$E_t^{en} = E_f^{en} + \sum_{n=6}^{N}[\langle E^{en} + P_n \rangle] + \sum_{n=1}^{N}[\langle E_I^{en} + P_n \rangle] \tag{5}$$

where E_t^{en} is the total enzyme concentration. By rearranging equation 4 to solve for $\sum_{n=1}^{N}\langle E_I^{en} + P_n \rangle$, and inserting it into equation 5, the free enzyme concentration can then be expressed as,

$$E_f^{en} = \frac{E_t^{en} - \sum_{n=6}^{N}[\langle E^{en} + P_n \rangle]}{1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I^{en}}} \tag{6}$$

Inserting equation 6 into equation 2 yields

$$K_{dis}^{en} = \frac{E_t^{en} - \sum_{n=6}^{N}\langle E^{en} + P_n\rangle}{1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I^{en}}} \times \frac{\phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}{\sum_{n=6}^{N}\langle E^{en} + P_n\rangle} \tag{7}$$

which can be rearranged into

$$\sum_{n=6}^{N}\langle E^{en} + P_n\rangle = \frac{E_t^{en}\phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}{K_{dis}^{en}\left(1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I^{en}}\right) + \phi^{en}\sum_{n=6}^{N}[(n-1)P_n]} \tag{8}$$

The total production of P_i from endopeptidase comprises of the hydrolysis reactions of the larger peptide chains P_k $(k = i+1, i+2, \ldots N)$ with each length k having the probability $\frac{2}{k-1}$ to result in a peptide chain of length i, as described in assumption 3. This probability, equation 3, and equation 8 then gives the net rate of reaction for P_i.

$$r_{P_i}^{en} = k^{en}\left(\sum_{k=i+1}^{N}\left[\frac{2}{k-1}\langle E^{en} + P_k\rangle\right] - \langle E^{en} + P_i\rangle\right)$$

$$= \frac{k^{en}E_t^{en}\phi^{en}(2\sum_{k=i+1}^{N}[P_k] - (i-1)P_i)}{K_{dis}^{en}\left(1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I^{en}}\right) + \phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}, \qquad i \geq 6 \tag{9}$$

Additionally, the net rate of reaction for smaller peptides is obtained as

$$r_{P_i}^{en} = k^{en}\left(\sum_{k=i+1}^{N}\left[\frac{2}{k-1}\langle E^{en} + P_k\rangle\right]\right)$$

$$= \frac{k^{en}E_t^{en}\phi^{en}(2\sum_{k=i+1}^{N}[P_k])}{K_{dis}^{en}\left(1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I}\right) + \phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}, \qquad 3 \leq i \leq 5 \tag{10}$$

in accordance with assumption 6.

4.3.2 Enzymatic Hydrolysis by Exopeptidase

The hydrolysis of external peptide bonds is also described by Michaelis-Menten kinetics and is represented below.

$$P_i + E_f^{ex} \overset{K_{dis}^{ex}}{\longleftrightarrow} \langle E^{ex} + P_i \rangle \overset{k^{ex}}{\longrightarrow} E_f^{ex} + P_{i-j} + P_j$$

$$P + E_f^{ex} \overset{K_I^{ex}}{\longleftrightarrow} \langle E_I^{ex} + P \rangle$$

In accordance with assumption 4, only the N-terminal is cleavable by the exopeptidase. Therefore, the substrate quantity is equal to the number of peptides with a degree of polymerization above or equal to six, i.e., $\sum_{n=6}^{N} P_n$. The dissociation constant is then defined as

$$K_{dis}^{ex} = \frac{E_f^{ex} \phi^{ex} P_i}{\langle E^{ex} + P_i \rangle} \cong \frac{E^{ex} \phi^{ex} \sum_{n=6}^{N} [P_n]}{\sum_{n=6}^{N} [\langle E^{ex} + P_n \rangle]} \tag{11}$$

The inhibition constant and enzyme mass balance for the exopeptidase is defined analogous to the endopeptidase. And similar to section 4.2.2, the net reaction rate is derived as,

$$r_{P_i}^{ex} = k^{ex} \left(\sum_{k=i+1}^{i+2} \left[\frac{1}{2} \langle E^{ex} + P_k \rangle \right] - \langle E^{ex} + P_i \rangle \right)$$

$$= \frac{k^{ex} E_t^{ex} \phi^{ex} (\sum_{k=i+1}^{i+2} [\frac{1}{2} P_k] - P_i)}{K_{dis}^{ex} \left(1 + \frac{\sum_{n=1}^{N} [P_n]}{K_I^{ex}} \right) + \phi^{ex} \sum_{n=6}^{N} [P_n]}, \qquad i \geq 6 \tag{12}$$

and for the smaller peptides

$$r_{P_i}^{ex} = k^{ex} \left(\sum_{k=i+1}^{i+2} \left[\frac{1}{2} \langle E^{ex} + P_k \rangle \right] \right)$$

$$= \frac{k^{ex} E_t^{ex} \phi^{ex} (\sum_{k=i+1}^{N} [\frac{1}{2} P_k])}{K_{dis}^{ex} \left(1 + \frac{\sum_{n=1}^{N} [P_n]}{K_I^{ex}} \right) + \phi^{ex} \sum_{n=6}^{N} [P_n]}, \qquad 1 \leq i \leq 2 \tag{13}$$

4.3.3 Bond Fractions

The variables ϕ^{en} and ϕ^{ex} are the fractions of hydrolysable bonds for endopeptidase and exopeptidase respectively. They are defined in equation 14 and 15,

$$\phi^{en} = \frac{\varepsilon^{en}\sum_{n=6}^{N}[(n-i)P_n]|_{t=0} - H^{en}(t)}{\sum_{n=6}^{N}[(n-1)P_n]|_{t=t}} \tag{14}$$

$$\phi^{ex} = \frac{\left(\varepsilon^{ex}\sum_{n=6}^{N}[(n-1)P_n]\big|_{t=0} - H^{ex}(t)\right)}{\sum_{n=6}^{N}[(n-1)P_n]\big|_{t=t}} \times \frac{2\sum_{n=6}^{N}P_n\big|_{t=t}}{\sum_{n=6}^{N}[(n-1)P_n]\big|_{t=t}} \tag{15}$$

where ε^{en} and ε^{ex} are the fraction of the initial bonds that are hydrolysable by endopeptidase and exopeptidase respectively. $H^{en}(t)$ is the amount of hydrolyzed internal peptide bonds and $H^{ex}(t)$ the amount of hydrolyzed terminal bonds at time t. The fraction for endopeptidase is assumed to be the remaining readily hydrolysable peptide bonds divided by the total amount of bond left at each time step. The fraction for exopeptidase is assumed as the product of the remaining readily hydrolysable peptide bonds fraction, times the fraction of peptide bonds within two steps of the N-terminal divided by the total amount of bonds.

4.3.4 Model Output

To summarize, the model is composed of N ordinary differential equations, each representing a chain length, which are described in equation 16.

$$\frac{dP_i}{dt} = r_{P_i}^{en} + r_{P_i}^{ex}, \qquad 1 \le i \le N \tag{16}$$

Additionally, for a total of $N + 2$ state variables, the "pseudo components" $H^{en}(t)$ and $H^{ex}(t)$ are also modeled as described in equation 17 and 18 respectively.

$$\frac{dH^{en}(t)}{dt} = \sum_{i=1}^{N}[r_{P_i}^{en}] \tag{17}$$

$$\frac{dH^{ex}(t)}{dt} = \sum_{i=1}^{N}[r_{P_i}^{ex}] \tag{18}$$

For the outputs of the model, the chain length state variables are lumped together into five components and described as a percentage of the total concentration according to table 4.1 and

equation 19. N is set to 113 for all simulations in this thesis and is motivated by the background theory in section 3.1 that the water-soluble protein has molecular weights mainly below 14 kDa. The Degree of Hydrolysis is calculated according to equation 20,

Table 4.1: The division into lumped components for the model output.

Lumped Compounds	Degree of Polymerization	Size (kDa)
A(t)	56-N	≥ 6.188
B(t)	37-55	4.0885 – 6.0775
C(t)	19-36	2.0995 – 3.978
D(t)	6-18	0.663 – 1.989
E(t)	1-5	≤ 0.5525

$$y_{Lumped\ Components} = \frac{\sum_{i=DP_{start}}^{DP_{end}} P_i}{\sum_{i=1}^{N} P_i} \times 100\%, \qquad Lumped\ Components = A, B, C, D, E \quad (19)$$

$$DH = \frac{\sum_{n=1}^{N}[P_n]}{(\frac{c_P}{M_{AA} - M_{H_2O}})} \times 100\% \quad (20)$$

where c_P is concentration of protein, $M_{AA}$is the average molecular weight of the amino acids in the protein, and M_{H_2O} is the molecular weight of water. M_{AA} is calculated by equation 21,

$$M_{AA} = \sum \left[F_{AA} \times M_{W,AA}\right] \quad (21)$$

where the average molecular weight is the sum of the products between the fraction of the amino acid, F_{AA}, and its molecular weight, $M_{W,AA}$. The fractions were calculated from table 3.1.

In total the model, and the system of ordinary differential equations (ODEs) can be described by the following form:

$$\frac{dx(t,\theta)}{dt} = f(x(t,\theta), u, \theta) \tag{22}$$

$$y(x, u, \theta) = g(x(t,\theta), u, \theta) \tag{23}$$

$$x(t_0) = x_0(\theta), \qquad t \in [t_0, t_{end}] \tag{24}$$

where x are the state variables, u is the model inputs (initial concentrations, temperature, pH) and θ is the fixed parameters. f describes the mathematical relations between the state variables and g is the mathematical function computing the measurable output variables y. x_0 are the initial values of the state variables and t is the time between t_0 and t_{end}. The explicit fourth order Runge-Kutta method within the SciPy package is used exclusively as the integration method for solving the differential equation systems in this thesis.

4.4 Synthetic Data Generation

The derived model was initially calibrated with two datasets (Rivero Pino et al., 2020). One set from hydrolysis of yellow mealworm protein by Alcalase, and the other by both Alcalase and Flavourzyme. The calibration was performed like the method described in next section but with a few modifications. The initial values of the chain length state-variables were estimated to get the best fit for the lumped compounds of the datasets. This was achieved by optimizing under the constraint described in equation 25,

$$y_0^A + y_0^B + y_0^C + y_0^D + y_0^E - S_0 = 0 \tag{25}$$

where y_0 is the initial concentrations of each lumped component and S_0 is the total protein concentration. Each lumped start concentration was evenly divided (in term of grams) among the chain length state variables of the lumped component and divided by the appropriate molar mass to give the initial values (in moles) for the simulations. The molar masses are calculated according to equation 26,

$$M_i = i \times \left(M_{AA} - M_{H_2O}\right) \tag{26}$$

where i is the peptide chain length and M_{AA} is calculated according to equation 21.

In total, the system of ODEs contains 115 differential equations, and 17 parameters were estimated. The parameter estimation was performed by minimizing the cost function (equation 27) in section 4.5. To handle both the bounds and the equality constraint in equation 25, Sequential Quadratic Programming was used for the minimization. More specifically, the "SLSQP" method in the minimize function of the SciPy package (Virtanen et al., 2020). The "goodness of fit" is evaluated by the root-mean-square deviation calculated by equation 28 in section 4.5.

After the parameter estimation, 9 synthetic datasets were generated according to a 2^{3-}factorial design with an added center point (see figure 4.1). The added center point can allow for better estimation of a curvature effect when RSM is utilized (Beg and Raza, 2021). The center point is the process conditions where the model was calibrated originally. The change in the parameters at other process conditions (temperature and pH) was conjectured based on the findings of Sousa Jr et al. (2004) and, Apar and Özbek (2010). The datasets were generated by simulating the model with the hypothesized parameters and appropriate concentrations. On top of the generated datapoints, artificial experimental noise was added from a gaussian distribution to make the data more realistically represent actual experimental data. The gaussian distributions mean was set to 0 and the standard deviation to the root-mean-squared deviation of the residuals from the parameter estimation.

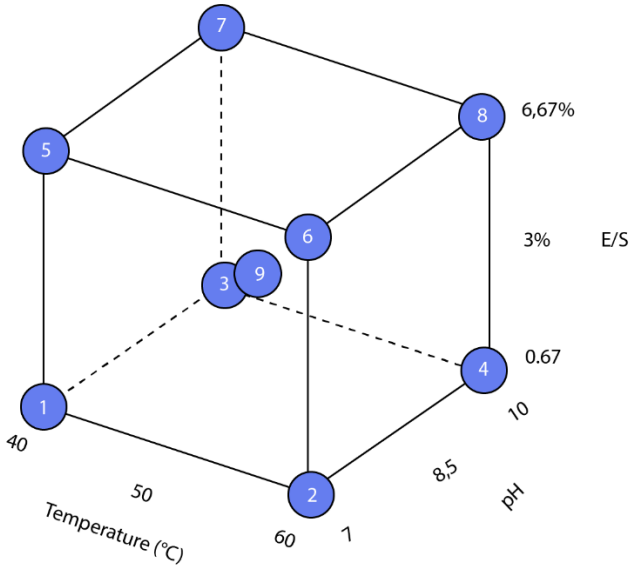

Figure 4.1: The 2^3-factorial design with an added center point used for synthetic data generation, showing the process conditions for temperature, pH, and the enzyme-substrate ratio.

4.5 Parameter Estimation

Parameters for the model were estimated through an optimization problem described below.

$$J(\theta) = \sum_{i=1}^{N_{DS}} \sum_{j=1}^{N_y} \sum_{k=1}^{N_t} \left[(y_{ijk}(x, u, \theta) - \widehat{y_{ijk}})^2 \right] \tag{27}$$

$$subjected\ to: \theta_{min} \leq \theta \leq \theta_{max}$$

The cost function, equation 27, is minimized to estimate the optimal values of the parameters θ. A common way to set up the cost function is the sum-of-squares, which is the case here. Where the errors between the experimental data and the model outputs are squared and summed together. In equation 27, $\hat{y}$ is the experimental value from dataset i, output j at time k. N_{DS} is the total number of datasets for the optimization, N_y the number of outputs for the model and N_t the evaluated time stamps. The optimization is bounded and the parameters each have an upper and lower limit during the minimization of $J(\theta)$. To solve the minimization problem, two

algorithms were tested. The Nelder-Mead, or Simplex, method is a popular direct search method which is efficient and does not rely on gradient evaluations. However, it does not guarantee a global minimum and can be very sensitive to the initial guesses of the parameters (Wang and Shoup, 2011). To evaluate a global minimum, the Differential Evolution algorithm (Storn and Price, 1997) was also tested. Both algorithms were deployed through the SciPy python package. The root-mean-square deviation was calculated for every optimization convergence according to equation 28 (Hyndman and Koehler, 2006). When the synthetic datasets were used for optimization, the system of ODEs contained 115 differential equations and 10 parameters were estimated.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{DS}} \sum_{j=1}^{N_y} \sum_{k=1}^{N_t} (y_{ijk}(x,u,\theta) - \widehat{y_{ijk}})^2}{N_{DS} \times N_y \times N_t}} \tag{28}$$

4.6 Identifiability of Parameters

Difficulties in estimating the parameters may arise if the model parameters display identifiability issues. Parameter identifiability is often divided into structural and practical identifiability. Structural identifiability refers to whether the parameters are uniquely determinable based on the formulation of the model and whether there are inherently indeterminable parameters. Practical identifiability issues arise from inadequate data used in the calibration. The main sources of practical non-identifiability are the parameters lack of influence on the observed output, and interdependence between the parameters (Gábor et al., 2017). One method to get an idea of the quality of the estimation is by computing the variance-covariance matrix and further, the correlation Matrix, of the parameters from the estimation. A linear approximation of the covariance matrix, $Cov(\theta)$, can be calculated according to equation 29 (Santacoloma, 2012).

$$Cov(\theta) = \frac{\min J(\theta)}{N-p} \left[\left(\frac{\partial y}{\partial \theta} \right)' \left(\frac{\partial y}{\partial \theta} \right) \right]^{-1} \tag{29}$$

$$Cor(\theta_i \theta_j) = \frac{Cov(\theta_i \theta_j)}{\sqrt{Cov(\theta_i \theta_i) * Cov(\theta_j \theta_j)}} \tag{30}$$

In equation 29, $\min J(\theta)$ is the minimized cost function, $N-p$ is the degrees of freedom (datapoints minus the number of parameters), and $\frac{\partial y}{\partial \theta}$ is the matrix of local sensitivities of the model outputs y to the parameters θ. In this thesis, they are calculated by numerical differentiation. Then, the correlation matrix, $Cor(\theta_i, \theta_j)$, is computed with equation 30. The values in the correlation matrix can vary between -1 and 1. Off-diagonal elements close to -1 or 1 indicate strong dependencies of parameter estimates. Essentially meaning that changing the parameter values can compensate for each other and have a canceling effect on the model, generating the same output. Hence, they would have poor identifiability (Brun et al., 2001).

4.7 Temperature and pH dependency with RSM

Following the discussion in section X, Response Surface Methodology was employed to describe the rate constants as functions of temperature and pH. The method follows the methodology described by Sarabia and Ortiz (2009).

A complete second order multivariate polynomial was evaluated as an empirical model (equation 31).

$$y = \beta_0 + \beta_1 * T + \beta_2 * pH + \beta_3 * T * pH + \beta_4 * T^2 + \beta_5 * pH^2 + \varepsilon \qquad (31)$$

In equation 31, y is the response variable, the betas are the coefficients and T and pH are the temperature and pH value respectively. ε is the error.

The model matrix is given below in eq 32 and is composed of the independent variables and their products according to the model function and the experimental design. Where the subscript represents the synthetic experimental datasets.

$$\boldsymbol{X} = \begin{bmatrix} 1 & T_1 & pH_1 & T_1 * pH_1 & T_1^2 & pH_1^2 \\ 1 & T_1 & pH_2 & T_2 * pH_2 & T_2^2 & pH_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_9 & pH_9 & T_9 * pH_9 & T_9^2 & pH_9^2 \end{bmatrix} \qquad (32)$$

With the definition of vector $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5\ ]'$, the equation system may be written as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (33)$$

where $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are vectors of the response values and errors respectively. Estimates of the coefficients by the least-squares method is then given by

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{34}$$

Estimated response values, $\widehat{\boldsymbol{Y}}$ can then be obtained by $\widehat{\boldsymbol{Y}} = \boldsymbol{\beta}\boldsymbol{X}$. With N as the number of response values (experimental datasets), and p the number of coefficients the variance of the errors is estimated as

$$s^2 = \frac{\sum[(\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^2]}{N - p} \tag{35}$$

5 Results and Discussion

5.1 Synthetic Data Generation

Initial efforts to calibrate the model yielded inadequate optimization results (data not shown) and the model was modified with adjusting parameters to allow for more flexibility. The insertion of the adjustment parameters ξ_A, ξ_B, ξ_C, and ξ_D into equation 10 can be seen in equation 36.

$$r_{P_i}^{en} = \frac{k^{en} E_t^{en} \phi^{en} (2\sum_{k=i+1}^{N}[P_k] - (i-1)\xi^j P_i)}{K_{dis}^{en}\left(1 + \frac{\sum_{n=1}^{N}[P_n]}{K_I^{en}}\right) + \phi^{en}\sum_{n=6}^{N}[(n-1)P_n]}, \qquad i \geq 6, \qquad j = A, B, C, D \qquad (36)$$

The four different parameters are applied according to the lumped components described in table 4.1. It is important to note here that even though no "strict" mass balances are defined within the unadjusted model, the probabilities of chain lengths forming from longer chains keep the mass balances intact, which was confirmed as the total protein mass remain in steady state throughout the simulations. However, introducing the adjustment parameters to get an acceptable fit, creates a mass balance issue that should not be ignored in further exploration of the model usage.

The estimated parameters used for generating synthetic data can be seen in table 5.1 along with optimization bounds and initial guesses. For the first dataset, the optimization has a root-mean-square deviation of 2.254873, the second dataset had a value of 2.205492 and the total value was 2.230319. The datapoints, along with the optimized solution can be seen in figure 5.1.

Table 5.1: Lower and upper bounds, initial guesses and estimated values of parameters using literature data.

Parameter	Lower Bound	Upper Bound	Initial Guess	Estimated Value
k^{en}	1e-5	1e3	3.85	3.71132
k^{ex}	1e-5	100	0.68	0.540613
K_{dis}^{en}	1e-2	1e5	365	366.347
K_{dis}^{ex}	1e-2	1e5	136	153.14
K_{I}^{en}	1e-3	500	203	203.783
K_{I}^{ex}	1e--3	500	0.54	11.0813
y_0^A	20	30	28.84	28.7115
y_0^B	0	1	1.478e-7	2.62874e-15
y_0^C	0	1	1.789e-7	2.64954e-14
y_0^D	0	10	1.15	1.15734
y_0^E	0	10	0.0078	0.131135
ξ^A	0	10	0.128	0.153993
ξ^B	0	10	6.59	6.38699
ξ^C	0	10	6.97	6.93443
ξ^D	0	10	9.58e-13	0. 386217
ε^{en}	0.1	0.5	0.2	0.1
ε^{ex}	0.1	0.5	0.2	0.1

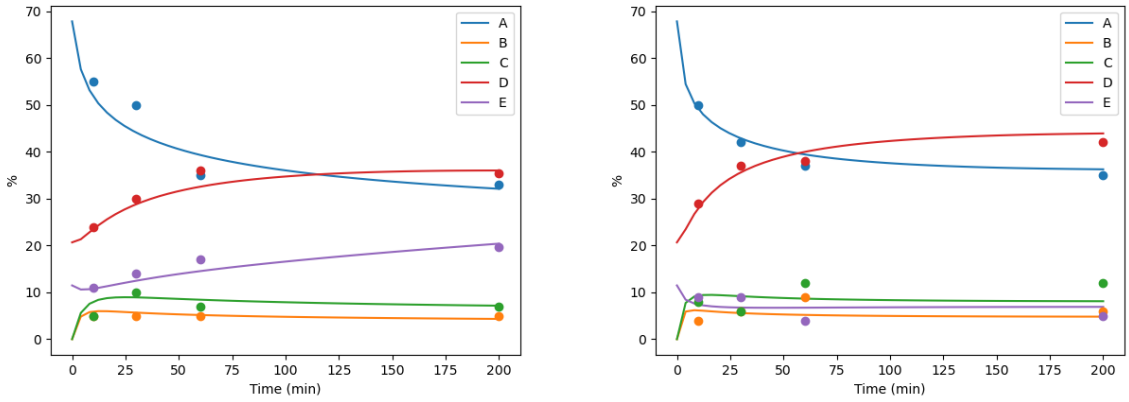

Figure 5.1: The plot on left is data and simulation results using both endo- and exopeptidase. On the right, only endopeptidase is used. Parameters from table 5.1 were used for the simulations.

5.2 Parameter Estimation

The Differential Evolution algorithm had to be discarded after initial optimization efforts. This was due to a lack of computational power and the algorithm required much more time than was reasonable regarding the time limitations of the project. All optimization results presented in this section are computed by the Nelder-Mead algorithm and a global minimum is therefore not guaranteed. For optimization of the synthetic datasets, the initial values were fixed according to the fractions defined by the parameters y_0^A, y_0^D, and y_0^E found in the previous section. The initial values for the state variables within the lumped components B and C were set to zero and the fraction of initially hydrolysable bonds, ε^{en} and ε^{ex}, were fixed according to table 5.1. The fact that the initial values used are computed through optimization is important to consider for future development of the model. The output will of course rely heavily on the initial state of the protein and experiments to determine the molecular weight distribution before enzymatic hydrolysis should be performed.

Calibration of the parameters were conducted both with each dataset individually, and in pairs where the temperature and pH are the same and only concentrations varied (see figure 4.1). Hence the kinetic parameters should have the same value for both datasets. The results of the estimations can be viewed in appendix A.

Examples of the model output are visualized below. Figure 5.2 show the complete molecular weight distribution and how it varies over time. The yellow bars are the initial values. The graph

is constructed from simulation results of the "purely" mechanistic model without the adjustment parameters ξ. Figure 5.2 is a simulation with only endopeptidase. Figure 5.3 is the result of a simulation with both endo- and exopeptidase present. It is clear from the figures that the model captures the behavior of the enzymes as the introduction of exopeptidase clearly promotes the release of dipeptides and free amino acids.

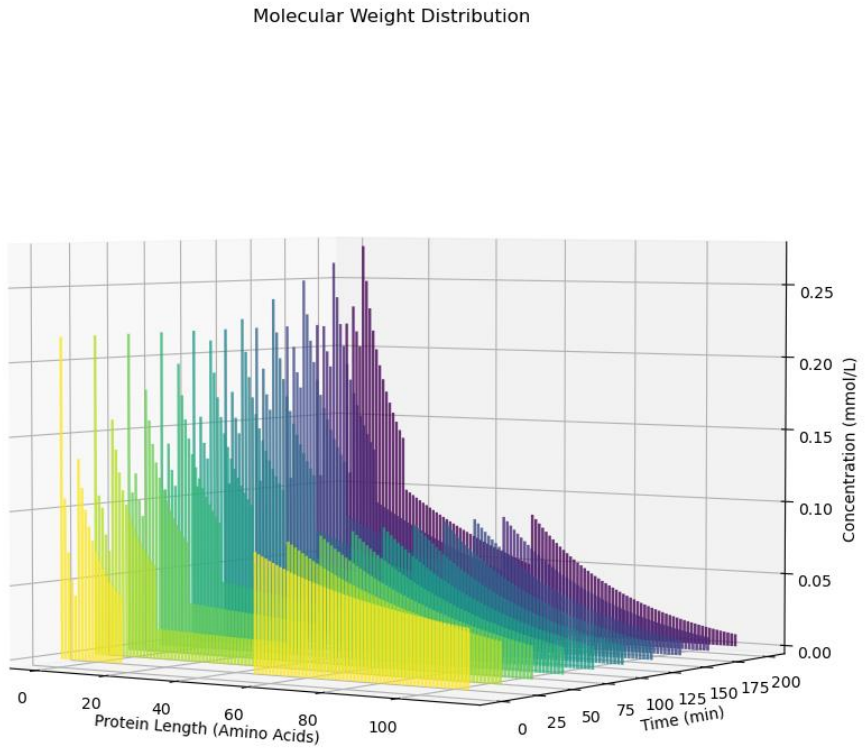

Figure 5.2: The molecular weight distribution of the protein over time using the un-adjusted model and only endopeptidase.

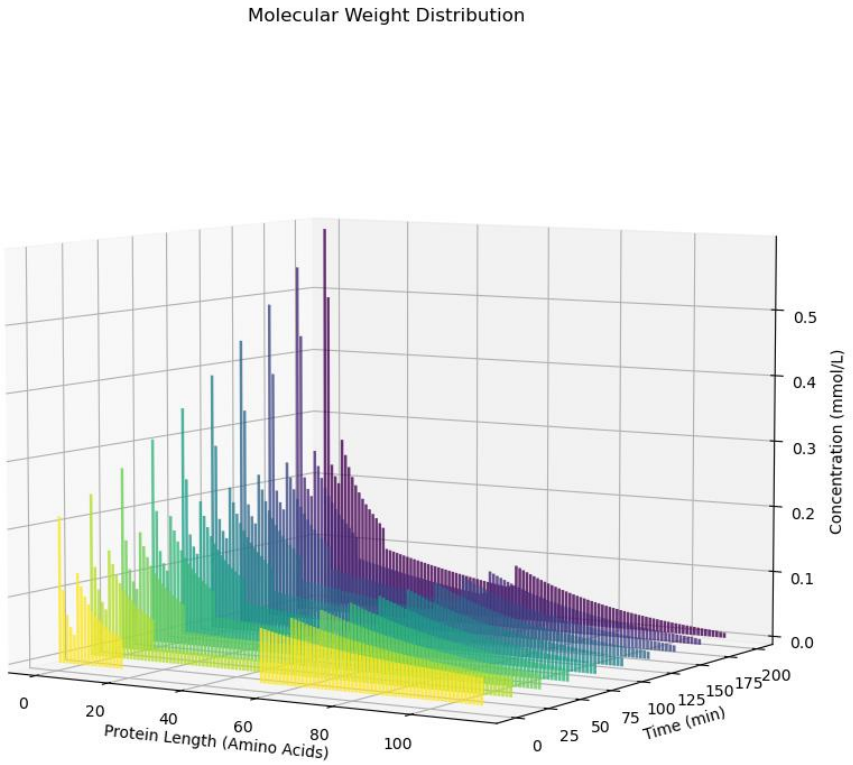

Figure 5.3: The molecular weight distribution of the protein over time using the un-adjusted model and both endo- and exopeptidase.

Figure 5.4 show the results from a simulation of the adjusted model calibrated with dataset 9 and its input variables (50 ℃, pH 8.5). Figure 5.4 (right) displays the lumped components as fractions of the total peptide concentration and is ultimately the model output that was compared to the data.

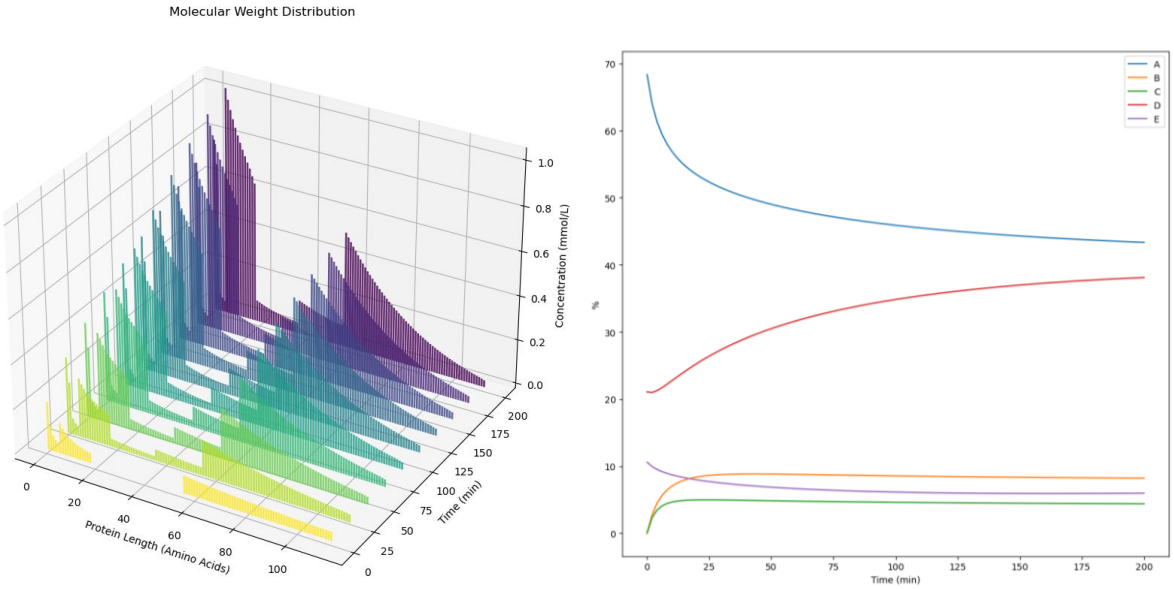

Figure 5.4: Outputs of the adjusted model calibrated with the synthetic dataset generated at temperature 50 °C and pH 8.5. Left: Molecular weight distribution over time. Right: Lumped components as percentages of the total peptide concentration.

The model structure appears to capture the general behavior of the process and optimizes with reasonable prediction errors (RMSD values ranging between 0,036 and 2,03). However, from the optimization results it clear that estimations are not reliable. The optimal parameter values vary drastically for datasets that should show similar results. Further, the rate constants for the enzymes does not generally capture the expected behavior of increasing with temperature and decreasing when operating at pH values above or below the optimal pH. This could be faulty due to the conjecture when the synthetic datasets were created and the stochastic nature of them. But it is more likely to stem from the Nelder-Mead algorithm converging in local minima and issues with identifiability of the parameters. This is discussed more in section 5.4.

5.3 RSM

The individual dataset calibrations of k^{en} and k^{ex} were used as response variables for the RSM method described in section 4.6. The calculated coefficients and error variance can be seen in table 5.2 and the plotted surfaces in figure 5.5.

Table 5.2: Table of estimated coefficients and error of variance for the empirical model in the RSM

Response	β_0	β_1	β_2	β_3	β_4	β_5	s^2
k^{en}	24,294	-4,595	22,172	0,030129	0,043362	-1,4346	3,075147
k^{ex}	4,3317	-0,79953	3,95686	0,004197	0,007266	-0,24227	0,6475311

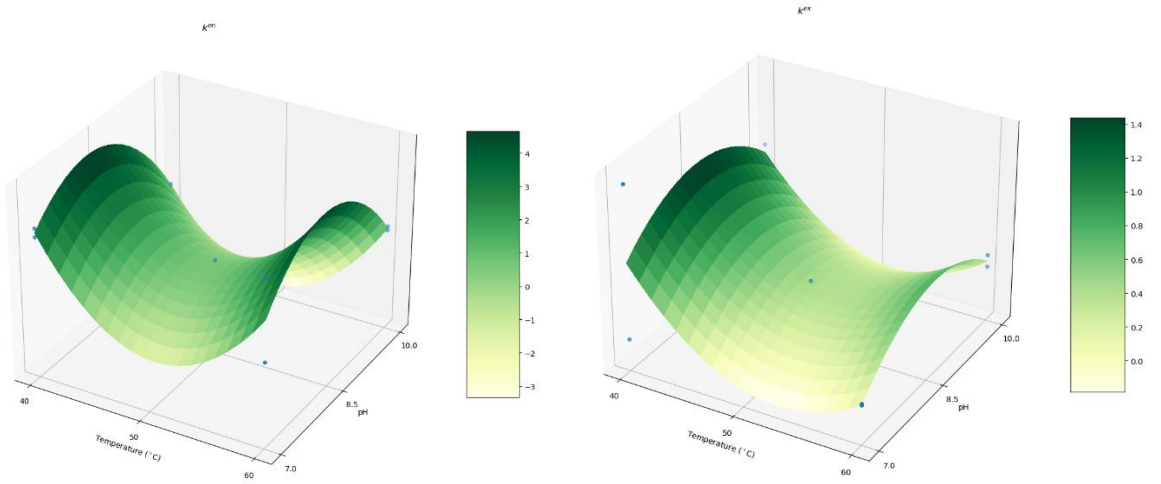

Figure 5.5: Surface plots of the estimated polynomials and the response variables used (blue dots) Left: k^{en}. Right: k^{ex}.

It is quite clear that the evaluated surface responses do not represent the expected behavior. Given the enzyme preparations temperature domain and the conditions in which the parameters are estimated, the parameters should only increase with the temperature variable. Since the highest temperature should not impose any loss in enzyme activity, which is further not included in the model structure. The "bell shape" of the pH dependency appears to be captured by the empirical model, however, the curve seems to be displaced when compared to what would be expected from real data and the enzyme preparations pH domains. The empirical model used would likely be able to capture the temperature and pH dependencies of the rate constants in a good way, given correct data. As discussed in the previous section, the parameter estimations do not appear to be very accurate, and hence, applying the response surface

methodology does not really contribute with any relevant information regarding pH and temperature dependencies.

5.4 Identifiability of the Parameters

As stated previously, the results of the parameter calibration indicate that there may be significant issues regarding the identifiability of the parameters. This is further motivated by the resulting correlation matrix of the estimation (see figure 5.6). As can be seen, all the absolute values in the off diagonals are unity (or very close to), which indicate poor identifiability among all parameters.

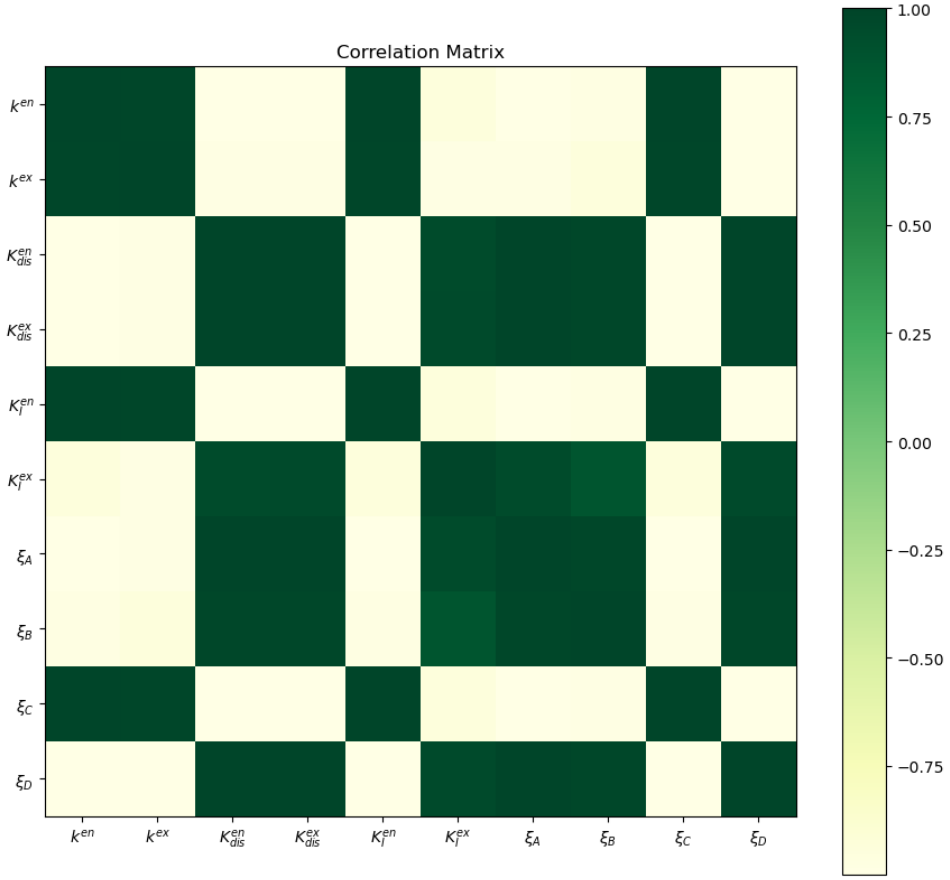

Figure 5.6: Correlation Matrix of the estimated parameters in the adjusted model.

This issue may arise from the fact that the data used for calibration is not the actual state-variables of the models. But the lumped components as fractions. This additional prediction

treatment hides the true nature of the parameters influence on the model predictions. And it is entirely possible that different simulation results and molecular weight distributions would give the same or very similar outputs when it is compared to the data. As a comparison, a correlation matrix was generated from an optimization where the data represented the actual state-variables of the system (see figure 5.7, left).

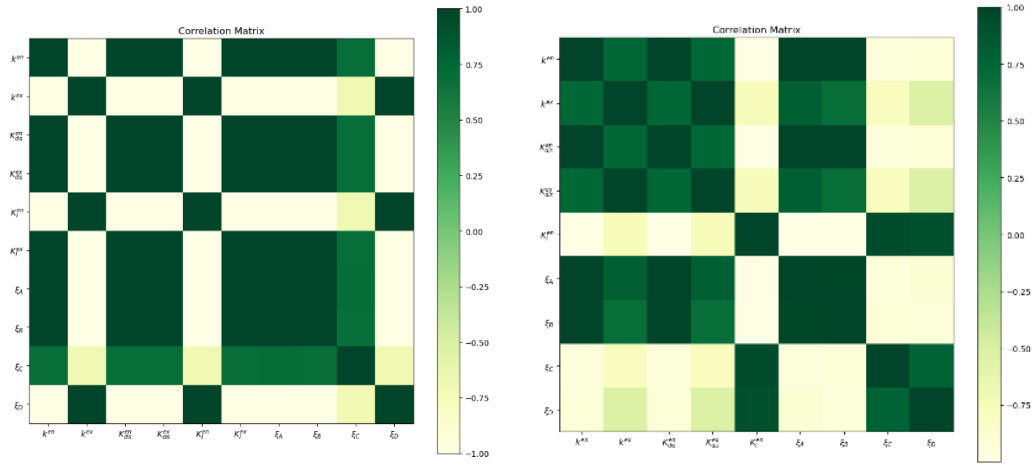

Figure 5.7: Correlation Matrices of the estimated parameters when data corresponding to the state-variables of the models were used. On the right, the parameter K_I^{ex} was removed from the model.

This generated a slight improvement regarding one of the adjustment parameters. The optimization also showed a significantly higher error on the estimation of the inhibition constant of the exopeptidase K_I^{ex} (calculated as $\sqrt{diag(Cov(\theta))}$) compared to the other parameters. This indicates the parameter has may not contribute to the model. A correlation matrix resulting from removing this parameter and again, optimizing with data representing the actual state-variables is shown in figure 5.7 (right). Again, a slight improvement is seen but the correlation coefficients are still very high, and it is difficult to determine a unique solution. Structural identifiability analysis may be beneficial to gain more knowledge whether it is the inherent model formulation, or the used data that causes the identifiability issues. Unfortunately, such analyses requires much computational work and are difficult to apply to large models (Gábor et al., 2017).

6 Conclusions and Future Work

To conclude this project, a mechanistic model was developed to describe the molecular weight distribution changes as a protein substrate undergoes enzymatic hydrolysis. The model showed good possibilities of capturing the expected behavior of the multi-enzyme system, but it is hard to draw definitive conclusions about its usefulness without obtaining real data on the system. An obvious drawback is the simplification of the protein structure and its generic repeating unit. Different amino acids and enzyme specificity will likely cause a hydrolysis that is not evenly distributed, which the model assumes. But it was a necessary assumption to keep the mathematical complexity at a reasonable level for this project. This is likely the reason that the un-adjusted model didn't have enough flexibility to get adequate calibration results. In addition, the assumed range for the molecular weight distribution, and the optimized initial conditions should be determined experimentally to create a more useful model in the future.

The parameter estimations and identifiability analysis revealed clear issues in determining unique solutions for the different process conditions. Improvement of the identifiability may be possible by not lumping together the state-variables before comparing to data. This does however increase the number of terms in the cost function and can come at a computational cost. But since the data for the lumped components would be derived from the full molecular weight distribution, it should not have a significant impact on experimental work that would be required. Prior to conducting experimental work, structural identifiability analysis might be wise to apply, though difficult, to determine whether there are inherently unidentifiable parameters.

Finally, the response surface method showed decent results that a complete second order multivariate polynomial could work as an empirical model to describe the temperature and pH dependencies of the rate constants. However, it is of course necessary that the responses used are correctly estimated.

7 References

APAR, D. K. & ÖZBEK, B. 2010. CORN GLUTEN HYDROLYSIS BY ALCALASE: KINETICS OF HYDROLYSIS. *Chemical Engineering Communications,* 197**,** 963-973.

ATKINS, P., JONES, L. & LAVERMAN, L. 2017. *Chemical Principles (International Edition) : The Quest for Insight,* New York, UNITED STATES, W. H. Freeman & Company.

BEAUBIER, S., FRAMBOISIER, X., FOURNIER, F., GALET, O. & KAPEL, R. 2021. A new approach for modelling and optimizing batch enzymatic proteolysis. *Chemical Engineering Journal,* 405**,** 126871.

BEG, S. & RAZA, K. 2021. Full Factorial and Fractional Factorial Design Applications in Pharmaceutical Product Development. Springer Singapore.

BRUN, R., REICHERT, P. & KÜNSCH, H. R. 2001. Practical identifiability analysis of large environmental simulation models. *Water Resources Research,* 37**,** 1015-1030.

CHEW, L. Y., TOH, G. T. & ISMAIL, A. 2019. Chapter 15 - Application of Proteases for the Production of Bioactive Peptides. *In:* KUDDUS, M. (ed.) *Enzymes in Food Biotechnology.* Academic Press.

COSTA, S., PEDRO, S., LOURENÇO, H., BATISTA, I., TEIXEIRA, B., BANDARRA, N. M., MURTA, D., NUNES, R. & PIRES, C. 2020. Evaluation of Tenebrio molitor larvae as an alternative food source. *NFS Journal,* 21**,** 57-64.

DE PRETTO, C., DE MIRANDA, L. C., DE SIQUEIRA, P. F., RIBEIRO, M. P. D. A., TARDIOLI, P. W., GIORDANO, R. D. C., GIORDANO, R. D. L. C. & COSTA, C. B. B. 2022. Mathematical modeling of enzymatic hydrolysis of soybean meal protein concentrate. *Chemical Engineering Communications,* 209**,** 338-350.

GÁBOR, A., VILLAVERDE, A. F. & BANGA, J. R. 2017. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Systems Biology,* 11.

GARCÍA ARTEAGA, V., APESTEGUI, M., MURANYI, I., EISNER, P. & SCHWEIGGERT-WEISZ, U. 2020. Effect of enzymatic hydrolysis on molecular weight distribution, techno-functional properties and sensory perception of pea protein isolates. *Innovative Food Science & Emerging Technologies,* 65**,** 102449.

GKINALI, A.-A., MATSAKIDOU, A. & PARASKEVOPOULOU, A. 2022a. Characterization of Tenebrio molitor Larvae Protein Preparations Obtained by Different Extraction Approaches. *Foods,* 11**,** 3852.

GKINALI, A.-A., MATSAKIDOU, A., VASILEIOU, E. & PARASKEVOPOULOU, A. 2022b. Potentiality of Tenebrio molitor larva-based ingredients for the food industry: A review. *Trends in Food Science & Technology,* 119**,** 495-507.

GRAYCAR, T. P., BOTT, R. R., POWER, S. D. & ESTELL, D. A. 2013. Chapter 693 - Subtilisins. *In:* RAWLINGS, N. D. & SALVESEN, G. (eds.) *Handbook of Proteolytic Enzymes (Third Edition).* Academic Press.

HABINSHUTI, I., NSENGUMUREMYI, D., MUHOZA, B., EBENEZER, F., YINKA AREGBE, A. & ANTOINE NDISANZE, M. 2023. Recent and novel processing technologies coupled with enzymatic hydrolysis to enhance the production of antioxidant peptides from food proteins: A review. *Food Chemistry,* 423**,** 136313.

HUIS, A., VAN ITTERBEECK, J., KLUNDER, H., MERTENS, E., HALLORAN, A., MUIR, G. & VANTOMME, P. 2013. *EDIBLE INSECTS future prospects fo food and feed security*.

HYNDMAN, R. J. & KOEHLER, A. B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting,* 22**,** 679-688.

KRISTOFFERSEN, K. A., AFSETH, N. K., BÖCKER, U., LINDBERG, D., DE VOGEL-VAN DEN BOSCH, H., RUUD, M. L. & WUBSHET, S. G. 2020. Average molecular weight, degree of hydrolysis and

dry-film FTIR fingerprint of milk protein hydrolysates: Intercorrelation and application in process monitoring. *Food Chemistry,* 310**,** 125800.

LANGE, K. W. & NAKAMURA, Y. 2021. Edible insects as future food: chances and challenges. *Journal of Future Foods,* 1**,** 38-46.

LENI, G., SOETEMANS, L., CALIGIANI, A., SFORZA, S. & BASTIAENS, L. 2020. Degree of Hydrolysis Affects the Techno-Functional Properties of Lesser Mealworm Protein Hydrolysates. *Foods,* 9.

MÁRQUEZ, M. C. & VÁZQUEZ, M. A. 1999. Modeling of enzymatic protein hydrolysis. *Process Biochemistry,* 35**,** 111-117.

MERZ, M., EISELE, T., BERENDS, P., APPEL, D., RABE, S., BLANK, I., STRESSLER, T. & FISCHER, L. 2015. Flavourzyme, an Enzyme Preparation with Industrial Relevance: Automated Nine-Step Purification and Partial Characterization of Eight Enzymes. *Journal of Agricultural and Food Chemistry,* 63**,** 5682-5693.

NIU, H., SHAH, N. & KONTORAVDI, C. 2016. Modelling of amorphous cellulose depolymerisation by cellulases, parametric studies and optimisation. *Biochemical Engineering Journal,* 105**,** 455-472.

PURSCHKE, B., MEINLSCHMIDT, P., HORN, C., RIEDER, O. & JÄGER, H. 2018. Improvement of techno-functional properties of edible insect protein from migratory locust by enzymatic hydrolysis. *European Food Research and Technology,* 244**,** 999-1013.

QI, W. & HE, Z. 2006. Enzymatic hydrolysis of protein: Mechanism and kinetic model. *Frontiers of Chemistry in China,* 1**,** 308-314.

RIVERO PINO, F., PÉREZ GÁLVEZ, R., ESPEJO CARPIO, F. J. & GUADIX, E. M. 2020. Evaluation of Tenebrio molitor protein as a source of peptides for modulating physiological processes. *Food & Function,* 11**,** 4376-4386.

RUAN, C.-Q., CHI, Y.-J. & ZHANG, R.-D. 2010. Kinetics of hydrolysis of egg white protein by pepsin. *Czech Journal of Food Sciences,* 28**,** 355-363.

SANTACOLOMA, P. D. G. A. 2012. *Multi-enzyme Process Modeling.* PhD, Technical University of Denmark.

SARABIA, L. A. & ORTIZ, M. C. 2009. 1.12 - Response Surface Methodology. *In:* BROWN, S. D., TAULER, R. & WALCZAK, B. (eds.) *Comprehensive Chemometrics.* Oxford: Elsevier.

SHI, D., HE, Z. & QI, W. 2005. Lumping kinetic study on the process of tryptic hydrolysis of bovine serum albumin. *Process Biochemistry,* 40**,** 1943-1949.

SOUSA JR, R., LOPES, G. P., TARDIOLI, P. W., GIORDANO, R. L. C., ALMEIDA, P. I. F. & GIORDANO, R. C. 2004. Kinetic model for whey protein hydrolysis by alcalase multipoint-immobilized on agarose gel particles. *Brazilian Journal of Chemical Engineering,* 21**,** 147-153.

STORN, R. & PRICE, K. 1997. *Journal of Global Optimization,* 11**,** 341-359.

TACIAS-PASCACIO, V. G., MORELLON-STERLING, R., SIAR, E.-H., TAVANO, O., BERENGUER-MURCIA, Á. & FERNANDEZ-LAFUENTE, R. 2020. Use of Alcalase in the production of bioactive peptides: A review. *International Journal of Biological Macromolecules,* 165**,** 2143-2196.

VALENCIA, P., ESPINOZA, K., CEBALLOS, A., PINTO, M. & ALMONACID, S. 2015. Novel modeling methodology for the characterization of enzymatic hydrolysis of proteins. *Process Biochemistry,* 50**,** 589-597.

VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE, E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P., VIJAYKUMAR, A., BARDELLI, A. P., ROTHBERG, A., HILBOLL, A., KLOECKNER, A., SCOPATZ, A., LEE, A., ROKEM, A., WOODS, C. N., FULTON, C., MASSON, C., HÄGGSTRÖM, C., FITZGERALD, C., NICHOLSON, D. A., HAGEN, D. R., PASECHNIK, D. V., OLIVETTI, E., MARTIN, E., WIESER, E., SILVA, F., LENDERS, F., WILHELM, F., YOUNG, G.,

PRICE, G. A., INGOLD, G.-L., ALLEN, G. E., LEE, G. R., AUDREN, H., PROBST, I., DIETRICH, J. P., SILTERRA, J., WEBBER, J. T., SLAVIČ, J., NOTHMAN, J., BUCHNER, J., KULICK, J., SCHÖNBERGER, J. L., DE MIRANDA CARDOSO, J. V., REIMER, J., HARRINGTON, J., RODRÍGUEZ, J. L. C., NUNEZ-IGLESIAS, J., KUCZYNSKI, J., TRITZ, K., THOMA, M., NEWVILLE, M., KÜMMERER, M., BOLINGBROKE, M., TARTRE, M., PAK, M., SMITH, N. J., NOWACZYK, N., SHEBANOV, N., PAVLYK, O., BRODTKORB, P. A., LEE, P., MCGIBBON, R. T., FELDBAUER, R., LEWIS, S., TYGIER, S., SIEVERT, S., VIGNA, S., PETERSON, S., MORE, S., PUDLIK, T., OSHIMA, T., et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods,* 17**,** 261-272.

WANG, P. C. & SHOUP, T. E. 2011. Parameter sensitivity study of the Nelder–Mead Simplex Method. *Advances in Engineering Software,* 42**,** 529-533.

ZHANG, Y. H. P. & LYND, L. R. 2006. A functionally based model for hydrolysis of cellulose by fungal cellulase. *Biotechnology and Bioengineering,* 94**,** 888-898.

ZHOU, C., YU, X., QIN, X., MA, H., YAGOUB, A. E. A. & HU, J. 2016. Hydrolysis of rapeseed meal protein under simulated duodenum digestion: Kinetic modeling and antioxidant activity. *LWT - Food Science and Technology,* 68**,** 523-531.

8 Appendix

Appendix A – Parameter Estimation Results

Parameter	Lower Bound	Upper Bound	Initial Guess	Dataset(s)		
				1	5	1&5
k^{en}	1,00E-05	10	3	3,44219	2,99431	0,279347
k^{ex}	1,00E-05	10	0,5	1,9561	0,001733	0,702868
K_{dis}^{en}	1,00E-05	inf	500	0,548991	898,945	8,07994
K_{dis}^{ex}	1,00E-05	inf	400	279,7	0,000227	823,077
K_{I}^{en}	1,00E-03	inf	200	0,003306	1833,52	1,98906
K_{I}^{ex}	1,00E-03	inf	20	26,5804	0,003627	46,181
ξ_A	0	inf	0,15	0,01695	0,130962	0
ξ_B	0	inf	7	7,16474	8,26311	7,71992
ξ_C	0	inf	7	4,8743	5,80905	5,64994
ξ_D	0	inf	0,4	1,02593	0,660251	0,77439

Parameter	Lower Bound	Upper Bound	Initial Guess	Dataset(s)		
				2	6	2&6
k^{en}	1,00E-05	10	3	0,125455	4,39179	0,205106
k^{ex}	1,00E-05	10	0,5	0,117135	0,102002	0,670008
K_{dis}^{en}	1,00E-05	inf	500	3,49E-05	308,165	0,110453
K_{dis}^{ex}	1,00E-05	inf	400	996,344	0,769963	510,13
K_I^{en}	1,00E-03	inf	200	2704,79	812,516	422,538
K_I^{ex}	1,00E-03	inf	20	0,023609	48,6128	26,752
ξ_A	0	inf	0,15	0,236119	0,14101	0,149282
ξ_B	0	inf	7	6,58642	12,4976	10,2932
ξ_C	0	inf	7	7,03851	6,72048	7,0007
ξ_D	0	inf	0,4	0,413815	0,296596	0,406112

Parameter	Lower Bound	Upper Bound	Initial Guess	Dataset(s)		
				3	7	3&7
k^{en}	1,00E-05	10	3	0,229087	0,143777	0,18609
k^{ex}	1,00E-05	10	0,5	1,10035	0,893649	0,604219
K_{dis}^{en}	1,00E-05	inf	500	0,55885	1,95004	5,30992
K_{dis}^{ex}	1,00E-05	inf	400	268,668	467,099	590,028
K_{I}^{en}	1,00E-03	inf	200	0,193837	374,522	922,494
K_{I}^{ex}	1,00E-03	inf	20	36,1845	8,88885	0,001
ξ_A	0	inf	0,15	0,020231	0,047633	0,037077
ξ_B	0	inf	7	3,47544	5,20137	4,8486
ξ_C	0	inf	7	16,7372	12,7084	13,4596
ξ_D	0	inf	0,4	0,338404	0,204584	0,225498

Parameter	Lower Bound	Upper Bound	Initial Guess	Dataset(s)			
				4	**8**	**4&8**	**9**
k^{en}	1,00E-05	10	3	1,13631	0,932848	0,627654	0,566107
k^{ex}	1,00E-05	10	0,5	0,454822	0,304221	0,052517	0,434693
K_{dis}^{en}	1,00E-05	inf	500	62,6442	35,1164	19,6373	10,3386
K_{dis}^{ex}	1,00E-05	inf	400	262,783	481,071	623,025	0,000533
K_{I}^{en}	1,00E-03	inf	200	267,36	497,187	367,315	20,9379
K_{I}^{ex}	1,00E-03	inf	20	6,56711	34,9805	38,8725	0,005762
ξ_A	0	inf	0,15	0,182023	0,138681	0,14266	0,036467
ξ_B	0	inf	7	6,8228	6,78684	6,69597	4,16981
ξ_C	0	inf	7	7,70274	7,67679	7,6703	15,9476
ξ_D	0	inf	0,4	0,503301	0,307866	0,330966	0,008594